Technologies Improving Access to Oral Histories: Fully Searchable Stories Presented in a Multimedia Web Portal

Michael G. Christel and Bryan S. Maher
Entertainment Technology Center
Carnegie Mellon University
Pittsburgh, PA 15213
1-412-268-7799
christel@cmu.edu

ABSTRACT

The Informedia research group at Carnegie Mellon University advances digital library research and the state of the practice of digital libraries by deploying existing, workable technologies into fielded operational video collections, evaluating that deployment with a focus on the human users, and iterating and refining the technology delivery to better suit the needs of the library patrons. The project has collaborated with The HistoryMakers African American oral history archive and used speech alignment, image processing, and language understanding technologies to promote multiple levels of access to fuel the viewing of the actual video recordings in a large oral history corpus. Since 2009, various versions of web interfaces were tested and refined with The HistoryMakers, the Harrisburg Pennsylvania Highmark Blue Shield Living Legacy Series (oral histories from Harrisburg), and in 2012 with the John Novak Digital Interview Collection: Experiences of the Civil Rights Movement oral histories established at Marygrove College. The design and development of the Adobe Flash application providing quick, easy access into digital video oral histories on the web will be discussed and demonstrated. This NSF-sponsored effort is available online at www.idvl.org, illustrated with multiple corpora showing that the work generalizes well. Tools will be overviewed, including a switch in 2012 to a different tool for automatically time-aligning segmented transcripts so that full-text searches mark where matches occur in the transcripts, along with automatically extracting named entities via OpenCalais so that map searches can be conducted based on locations mentioned in oral histories. These tools and interface are freely available to the oral history community. This talk will showcase the benefits for improving access down to the individual story level through synchronized metadata.

1. INTRODUCTION

The Informedia research group at Carnegie Mellon University has worked with oral history archives to apply automatic speech alignment, image and language processing in order to generate time-aligned metadata for use in accessing the video narratives. A study in 2007 showed the value of representing the oral histories in video form, especially for exploratory search [1]. Oral historians have welcomed the utility of such tools to provide direct access to the audio and video content of their collections [3, 6]. They bemoaned the prior state of the field in which text transcripts were the sole accessible representation, with original audio and video sources either set aside or even discarded altogether [6].

This paper discusses the development of an oral history interface deployed at www.idvl.org showcasing three video collections:

1. The ScienceMakers: African Americans and Scientific Innovation oral history collection (later referred to simply as "ScienceMakers"), a National Science Foundation (NSF) funded collection also originating from and curated by The HistoryMakers non-profit organization based in Chicago, IL.
2. The Highmark Blue Shield Living Legacy Series, Harrisburg, PA (later referred to as "Harrisburg"), recording the memories of 150 Harrisburg area residents in celebration of the city of Harrisburg's 150 years of incorporation in 2010.
3. Portions of the John Novak Digital Interview Collection: Experiences of the Civil Rights Movement series from Marygrove College [12].

An additional archive has been processed with the same technologies and tools discussed here: The HistoryMakers Digital Archive, a robust collection of nearly 1000 hours of African American life oral histories segmented into over 16,000 stories. The HistoryMakers Digital Archive has been moved to its own web site at http://www.thehistorymakers.com which requires an account from that web site. Lessons learned in moving processing capabilities and web hosting out from the Carnegie Mellon research team at www.idvl.org to thehistorymakers.com will also be overviewed here.

The processing, schemas, and interface code showcased at www.idvl.org with these three corpora is being shared as open source so that others can derive the benefits and tailor the interfaces reported here. Use of these three oral history corpora will continue to be monitored to improve the interfaces into oral histories, with the ScienceMakers likely to move into thehistorymakers.com for continuity with their collections. The addition of the audio-only Experiences of the Civil Rights Movement series is noteworthy in that it demonstrates the applicability of the Informedia digital video processing tools and framework to an audio-only corpus. This is not surprising, in that other work with audio-centric or exclusive corpora such as the National Gallery of the Spoken Word [5] or the Czech historical radio archive [10] has underscored the value of speech-aligned transcripts for subsequent search and exploration.

Generating transcripts automatically through speech recognition is possible but with a huge error range and set of specific research issues for dealing with noise, accents, and vocabulary, with additional challenges in utilizing natural language processing techniques against such materials [3]. The HistoryMakers, Harrisburg, ScienceMakers, and Marygrove College corpora benefit from Informedia processing for automated speech alignment of provided transcript text (rather than generation of such text from the speech signal). Additional Informedia processing includes a Lemur text search service [7] against the

text metadata and geographic search based on automated extraction of locations from text through named entity extraction.

The first iteration of what is now posted at idvl.org began with The HistoryMakers and a stand-alone Windows application deployed at numerous universities and organizations. Synchronized transcripts were a major feature of the work since before 2006 and praised by users in early controlled studies [1, 2]. Text-based video navigation through synchronized transcripts is an efficient means for browsing oral histories, with de Jong noting additional value that NLP can bring to synchronized metadata for oral history collections [3]. The first iteration concluded with clear indications from surveys and transaction logs that users were frustrated with the closed delivery infrastructure and wanted a better delivery architecture offering 24/7 access to the oral histories from their own computers [2]. The second iteration in early 2010 resulted in fielding a Flash application on an openly accessible website, idvl.org. During the second iteration the Harrisburg corpus was also addressed in order to test the generality of the methods used. After six months of use (July-December 2010), the Flash interface was revised. The revisions were based on transaction log data, comments volunteered by users through email and through a comments interface in the website, novice and expert commentary at workshops and through OHA demonstration sessions, and careful review of literature discussing facets and exploratory search (e.g., [8, 13]). A presentation at OHA 2011 emphasized the interface changes and their underlying rationale, with the ScienceMakers deployed using the same interfaces and processing tools in July 2011. In 2012, the Marygrove collection was added. This paper walks through the processing steps used to deliver the interface experience present for these oral history sets at www.idvl.org.

2. PROCESSING STEPS

The archivist initiates the creation of the web portal into their collection by providing a plain text mark-up document with segmented transcripts for a video interview, and additional provenance and other categorical detail regarding the interviewee, interviewer, and interview. The input document contains both special "mark-up" lines and content lines, with the mark-up being a small set of tags indicating the provenance information, story titles, story start and end times, and story transcripts. The categories, i.e., facets, may be specialized for each collection, e.g., ScienceMakers offer a “Maker” category like PoliticalMaker, while all three of Harrisburg, ScienceMakers, and Marygrove offer gender, birth year, and O*NET job types. Transcription and segmentation of the interviews was left with the human archivist rather than automated to give great control over the quality of the transcript, the granularity of the story length, and the tone of the story titles. Subsequent tedious steps to provide a time-aligned transcript are automated, described below according to purpose, function (source), and requirements.

2.1 User-Directed Story Segmentation

An automated tool checks the input document for completeness (e.g., some tags like interviewee full name for interface labels and last name for sorting are required), and correctness (a configuration file indicates the minimum and maximum acceptable story lengths for the corpus). The user can also mark some transcript text as unspoken via square brackets, e.g., perhaps an aside to disambiguate a city to a particular state as in "I really enjoyed my time in Rochester [New York]." A simple automated check for correctness is that all square brackets are balanced to properly delineate the spoken words from the unspoken words.

If all input is correct, then XML output files are produced which will trigger subsequent Informedia processing by Windows scripts that will result in content being posted for the oral histories into a Microsoft SQL Server MDF database file. The Microsoft platform was chosen for processing scripts based on both availability of the platform to our oral historian partners, and the good performance that the Microsoft automatic speech recognition engine provides for the transcription alignment step. The XML files from this step move into both speech alignment processing and additional automated text processing.

2.2 Speech Alignment Processing

Through Windows Speech Recognition (WSR) present with the Microsoft Windows 7 operating system, a pseudo-transcript is automatically generated from each story segment's soundtrack. The pseudo-transcript likely contains errors, but is suitable for matching against the human-generated transcript to produce a list of high-confidence match words. The transcript character offset and time spoken for each match word is stored in the database to then automatically time-align the transcript words to the segment's soundtrack. The alignment data is used to provide quick access to the relevant story section for text searches as well as "bouncing ball"-style transcript highlighting in the interface.

2.3 Web Video Segmentation and Keyframe Extraction

In parallel with alignment, the source video is broken up into smaller story-level video segments for use with the idvl.org Flash client application. The open source software FFmpeg [4] is used to transcode subsections of the source video into smaller web-ready MPEG-4 files. The resulting files are MPEG-4 AVC video with AAC audio stored in an MP4 container. The MP4 video is then optimized for HTTP streaming using the public domain tool "qt-faststart" [9].

A representative frame from each segment video is stored in the database. For corpora like Marygrove College where the source interview files are audio-only, a representative image for the story can be used as a visual reference, again using FFmpeg for convenience. Specifically, FFmpeg extracts a representative frame from the story segment video and writes it to a temporary file as a PNG image. The resulting image is read into memory, converted to JPEG, then stored in the database. All temporary files are cleaned up prior to completing the task.

2.4 Additional Automated Text Processing

Geocoding is run to identify common place names within the transcript and tag them with their geo-spatial coordinates. GeoCoding is provided by the OpenCalais web service from Thomson Reuters [11]. Transcript text is sent over the internet to the OpenCalais web service which analyzes the text and returns a list of entities. The geographical entities are extracted from the results and added to the database. The end-user must apply for a free API key to access the OpenCalais web service. The API key must be added to a local configuration file to activate the GeoCoding functionality.

OpenCalais is limited to 50,000 transactions per day at a maximum rate of 4 transactions per second. Processing one story segment's transcript equates to one OpenCalais transaction, therefore, there is a limit of 50,000 story segments per day at a

rate not exceeding 4 story segments per second. In practice, the comparatively slow speed of processing (order of several minutes per segment) would prevent this limit from ever being reached.

2.5 Searchable Index Generation

The story titles and transcripts are indexed using the Indri search engine of the Lemur toolkit [7]. The Flash interface presents an easy way to use Indri "and" search (search across all words), "or" search (search for matches against any of the given words), and adjacency search (to search phrases). This step produces a set of index files that must be deployed on the web server where the web service is hosted, the service that connects the Flash application to the MDF database. The web service is written in C# using ASP.Net and Microsoft Internet Information Services (IIS) for the ease of integrating with Microsoft SQL Server in order to connect to the MDF database file. Hence, the web server for hosting corpora using these described processing steps is an IIS web server (as used for idvl.org).

2.6 Deployment to a Web Site

For ScienceMakers and Harrisburg, individual interviewees and segment story titles listed in html pages act as a web-indexable table of contents on idvl.org, and link into the Flash application to show just those story segments. For example, "Ho-Thanh Nguyen" is discussed at the following web page which provides links to each of her stories as well as one link for all 15 stories, http://www.idvl.org/harrisburglivinglegacy/Bio120.html. For the Marygrove collection, a separate web site lists more information about the archive, with the idvl.org site containing only the searchable archive by choice. This illustrates the flexibility afforded to the oral history archive holder: allow deep web search index access into the archive (e.g., allow Google, Bing, and other indexing to full transcripts), allow shallow indexing (as is done with ScienceMakers and Harrisburg, allowing indexing of interview abstracts but not indexing of full transcripts), or allow no special extra indexing (as is done with the Marygrove collection). Regardless of the level of extra content exposed in additional automatically generated web pages from the MDF data store for the oral history corpus, the full contents of the oral history set remain accessible through the Flash front-end client.

3. WEB INTERFACE TO THE ORAL HISTORIES

At OHA 2011, the following lessons were shared regarding the Flash front-end interface to the oral histories:

- Simpler navigation and bookmarking: Through open source projects *swfaddress* (to provide deep linking for Flash), *swfobject* (to embed Flash in html), and *swffit* (to resize Flash with browser window), the idvl.org portal to oral histories allows users the convenience of bookmarking video sets and individual stories, promoting sharing across the historical research and lifelong learning community.

- Better facet communication: following advice that faceted interfaces for digital libraries present facets along with the data, rather than in a separate control page [8], the faceted search panel is shown along with the table of contents, and with the video result set.

- Improved information seeking: the many ways to seek information are listed on the first-seen "main" state of the application (text search, map search, table of contents, saved sets), with "berrypicking" supported through buttons and drag and drop. Berrypicking is the process of gathering a bit of information at a time through evolving and complementary searches [13], and is supported in the interface with a Play List area. The original interface has a "Collection" area where users had to drag and drop items, an interface paradigm that was unfamiliar to some users. The new interface features prominent green plus and red minus buttons to add and remove items from your "Play List." Through bookmarking, you can share your Play List with others.

One goal of the oral history portal is to promote access to the primary source materials: the audio and video content noted for its value to the archive's patrons [6]. The fear that segmenting the interview into story units will weaken the corpus because segmentation can be arbitrary [6] is mitigated by allowing the user to seamlessly chain temporally from one story to the next adjacent video sequence using the buttons for that purpose presented in the video page view. Hence, information seeking can be following a temporal chain, initiating an analytic search strategy, or serendipitous browsing of facets provided in the interface.

3.1 Corpus-Specific Styling

Through Flex skins and css files defining the look for Flex components, Adobe Flex Builder allows for the generation of .swf (Flash) files presenting different visual themes: rose for ScienceMakers, gold with blue accents for Harrisburg, green for the Marygrove collection. Linked in logo images from the web site allow for further tailoring of the interface. Such customization is part of the open source resources demonstrated and made available through idvl.org. The front door to the Marygrove collection at idvl.org is shown in Figure 1.

Figure 1. Front door to the John Novak Digital Interview Collection: Civil Rights Movement Series at idvl.org/JohnNovak, with a means to browse table of contents, review an assembled Play List, do a full-content text search across story titles and transcripts, or do a map search.

Given its frequent access, the video page is of primary importance, with the time-aligned transcript allowing for fast access to neighborhoods of interest within even very long oral history stories. The time-aligned transcript makes the video more accessible [2] and offers potential to keep users on-site longer in support of exploratory search [1]. Figures 2, 3, and 4 show a sequence of activity via three screen shots: a map search for "North Carolina" using the Google Maps interface, the results of

that query, and a Esther Terry story that matched because it mentioned Greensboro among other locations (and via the disambiguation and gazetteer information from OpenCalais, Greensboro is properly geocoded to "Greensboro, NC" in this context and hence is color-coded (yellow) in the transcript display as a match to a "North Carolina (NC)" search).

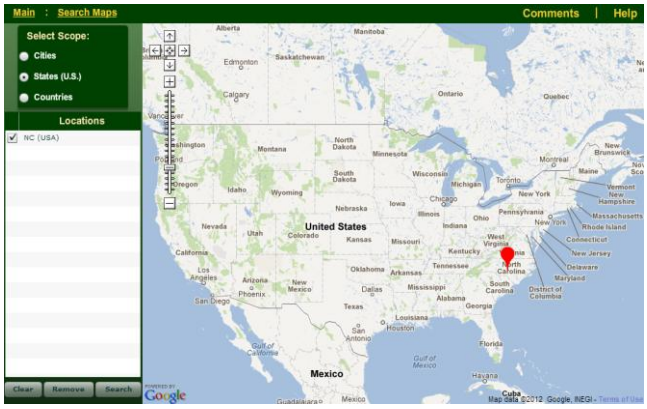

Figure 2. Google Maps interface in support of map search by city, state, and/or country using the metadata produced by the OpenCalais geocoding processing step.

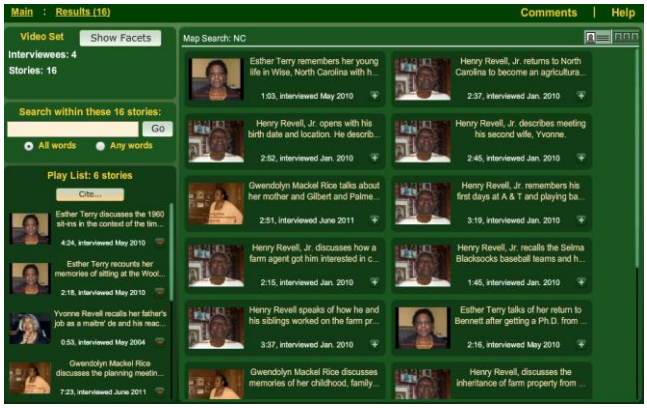

Figure 3. Results from North Carolina map search, along with a user-assembled Play List of notable stories at lower left.

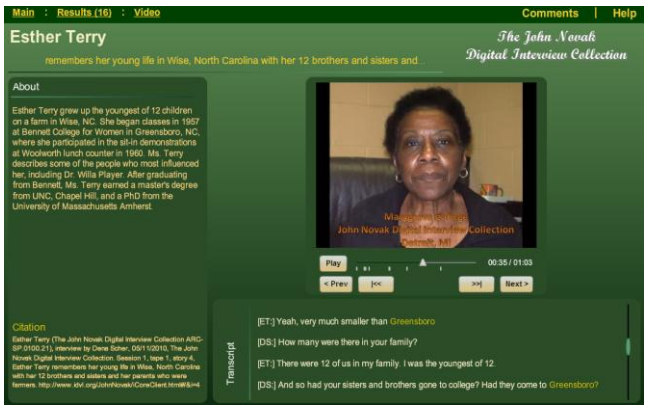

Figure 4. Video details page for 1 story from Fig. 3 results.

The locations of matching areas following a query are represented on the video timeline, with button access to quickly jump forward and back through match areas. The proper way to cite the page along with a link is given in the lower left, so that a particular story can be referenced and returned to with ease.

Citing the play list also became an area of concern: after using the system and berrypicking results, how do you easily cite the assembled Play List? For example, a student interested in women in science might search on these terms, review stories, and start assembling a Play List as shown in the lower left of Figure 5 for the ScienceMakers collection. He or she adds stories by clicking the green "+" icon on relevant stories. She can click the "Cite..." button at any time to get a text list of the citations, along with a link to the FAQ help page for further context.

Figure 5. Zoomed-in view of Play List area (3 items) with "Cite..." button to provide text list (3 citations) for easy reuse.

To tie together the processing steps of the prior section with the interface, consider the video details page from the Harrisburg collection following a query on Olympics, shown in Figure 6. User-provided data gives the collection name "Rusty Owens," the abstract, the transcript, the story title, and other provenance information used to formulate the citation text shown in lower left. The MP4 video plays in the central area with a timeline beneath it allowing for quick seeking into the video to match points where the word "Olympics" is mentioned. The time alignment processing provides the synchronized metadata in the MDF data store supporting this interface capability. Story sequencing from the original input file and timing is retained so the user can follow a temporal chain in the interview, from one story to the next. The Indri search engine indicates which words matched in the transcript so that they can be color-coded.

Figure 6. Video details page, calling out some portions of the interface.

4. CONCLUSIONS AND FUTURE WORK

Placing the oral history collections on the web allows discoveries of the materials by not only dedicated researchers, but also casual browsers. A Flash interface allows for visual richness and cross browser compatibility, is a prevalent way of displaying video on the web, and through open source tools can support expected browser button navigation and bookmarking as well to internal Flash program states. These features are now in active use for three online oral history collections at www.idvl.org.

After working with The HistoryMakers for both their full Digital Archive and the ScienceMakers, with the Harrisburg and Marygrove College collections, the processing steps documented here represent a set of simple yet robust technologies that open up oral history corpora for greater access by a broad group of internet users. The MDF database schema is rich enough to support interviews across multiple sessions and supports customizable facets, yet remains compact enough to support an easy learning curve and broader adoption. The interface, however, will likely need upgrading in that Flash support is dwindling on mobile devices, yet Internet use is expanding rapidly into such devices, including tablets and smart phones. A likely next step for this work is to port the interface to HTML5 so that it can run across a broader set of platforms, providing even more access into the underlying oral histories.

For OHA 2011, our work concluded with a question on exploratory search: can we produce an interface that encourages staying within the Flash application to explore a broader tapestry of stories? Web exploratory search is discussed well in [13] and fits the hopes of oral historians to have their source audiovisual forms explored more deeply [6]. This emphasis is natural for oral history web portals: they cannot compete on the basis of quick fact-finding (for that, text transcript-only representations work fine, and authoritative text sources may exist as well), but they do offer unique opportunities for experiencing and exploring first-person multimedia narratives. Will the resources available at www.idvl.org encourage not only a dive into a story to satisfy an information need, but also a more thorough experience with the corpus in which additional stories are reviewed, played, and shared? Such questions will motivate our analysis of Informedia technologies applied to oral history libraries, and drive future work that may carry lessons learned from our Flash application into the development of new HTML5 portals.

5. ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under grants IIS-0705491 and DRL-0917612. The work of ETC graduate students Andy Korzik, Xiaoxi Liu, and Srinavin Nair is featured in the discussed Flash interfaces. ScienceMakers work is aided greatly by the skills of Alison Bruzek, Marta Grabowski, Paul Mackey, and Dan Johnson, with ScienceMakers and HistoryMakers work made possible via the active collaboration of the HistoryMakers executive director Julieanna Richardson. Collaboration with Dena Scher led to the establishment of the Civil Rights Movement collection at http://www.idvl.org/JohnNovak.

6. REFERENCES

[1] Christel, M. and Frisch, M. 2008. Evaluating the contributions of video representation for a life oral history collection. In *Proc. ACM/IEEE-CS Joint Conf. Digital Libraries* (Pittsburgh, PA, June 2008). ACM, New York, NY, USA, 241-250..

[2] Christel, M., Stevens, S., Maher, B., and Richardson, J. 2010. Enhanced exploration of oral history archives through processed video and synchronized text transcripts. In *Proc. Conf. on Multimedia* (Florence, Italy, Oct. 2010). MM '10. ACM, New York, NY, 1333-1342. DOI=http://doi.acm.org/10.1145/1873951.1874215.

[3] de Jong, F. NLP and the humanities: the revival of an old liaison. 2009. In *Proc. Conf. European Chapter of the Association for Computational Linguistics* (Athens, Greece, March/April 2009). Association for Computational Linguistics, Stroudsburg, PA, USA, 10-15.

[4] FFmpeg, http://ffmpeg.org/ (accessed 10/2012).

[5] Hansen, J.H.L., Huang, R., Zhou, B., Seadle, M.S., Deller, J.R., Gurijala, A., Kurimo, M., Angkititrakul, P. SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word. In *Proceedings of IEEE Transactions on Speech and Audio Processing*, 2005, 712-730.

[6] High, S., and Sworn, D. 2009. After the Interview: The Interpretive Challenges of Oral History Video Indexing. *Digital Studies / Le Champ NuméRique* 1, 2.

[7] *Lemur Toolkit for Language Modeling and Information Retrieval*. http://www.lemurproject.org (accessed 10/2012).

[8] Medynskiy, Y., Dontcheva, M., and Drucker, S. 2009. Exploring websites through contextual facets. In *Proc. Conf. on Human Factors in Computing Systems* (Boston, MA, April 2009). ACM, New York, NY, USA, 2013-2022.

[9] Melanson, M. qt-faststart, listed with FFmpeg at http://ffmpeg.org/doxygen/trunk/qt-faststart_8c-source.html (accessed 10/2012).

[10] Nouza, J., et al. Making Czech Historical Radio Archive Accessible and Searchable for Wide Public. *Journal of Multimedia* 7(2), April 2012, 159-169.

[11] OpenCalais, powered by Thomson Reuters. http://www.opencalais.com/ (accessed 10/2012).

[12] Scher, D., and Malmsten, C. The John Novak Digital Interview Collection: Experiences of the Civil Rights Movement series, Marygrove College (accessed 10/2012), //research.marygrove.edu/novakinterviews/civil_rights.html.

[13] Wilson, M., Kules, B., schraefel, m. c., and Shneiderman, B. 2010. From Keyword Search to Exploration: Designing Future Search Interfaces for the Web. *Foundations and Trends in Web Science* 2, 1 (January 2010), 1-97.